Exploring the relationship between RNA-mediated gene duplication and linkage modification

Johnathan Lo

12/1/20

1 Abstract

Retrogenes have been intensely researched since their discovery several decades ago. They are of particular interest in evolutionary biology due to their role in phenomena like the out-of-the-X pattern or retrogene replacement. This interest also stems from the processes that participate in their fixation, which range from subfunctionalization to enhanced rates of evolution due to relaxed evolutionary constraints. One process that has not been investigated is the influence of selection on linkage in manipulating the fixation of retrogenes. Genes are often subject to epistatic interactions where certain alleles work well together and others do not; as such, selection can act on combinations of alleles, and the frequency of different combinations is dependent on linkage. When new alleles emerge at one locus, they can potentially alter the selective forces on linkage with other loci, thereby driving retrogenization. This type of process can serve as a complement to modifier allele models of linkage modification, by invoking a physical rearrangement of genetic material rather than an increase in recombination between loci. The potential interaction between selection on linkage and retrogene fixation can manifest in two ways: either when the fixation pattern of retrogenes is influenced by selection on linkage, or when selection on linkage is resolved through increased rate of fixation. In this paper, I outline this proposal and the steps I will take to test my hypotheses.

2 Introduction to RMGD

RNA-mediated gene duplications (RMGDs) were first described in the early 1980s[1]. They emerge as a byproduct of reverse transcriptase (RT) activity, which itself is encoded and produced by Class I TEs in the genome. RMGDs are created when the RT enzyme interacts with an mRNA to mediate its insertion into the genome[1,2]. This process can be viewed as a counterpart, or alternative, to the more widely-studied DNA-mediated gene duplication (DMGD). Because RMGD tends to omit the regulatory elements and other context-specific sequences that help regulate expression, these retroposed gene copies were originally assumed to be non-functional.

However, as research has progressed, an increasing number of functional RMGDs have been found, termed ”retrogenes”. These have been characterized with diverse roles, ranging from spermatogenesis to courtship behavior[1]. Non-protein coding retrogenes have been shown to make some contribution to gene regulation through production of fragmentary peptides and siRNAs, including the famous TP53 duplications in elephants[2]. From an evolutionary perspective, RMGD allows genes to breach a wider evolutionary space by removing constraints like intron-exon junctions and regulatory sequences[1], and contributes to phenomena like exon shuffling and protein chimerism[1,2]. Other retrogenes have been found to contribute to antiviral defenses, novel phenotypes in hormone-pheromone metabolism, brain, and courtship behavior[3,4,5,6]. Retrogenes also sometimes supplant the function of parental genes (so-called ”orphan” retrogenes)[7]. Non-protein coding retrogenes have been shown to make some contribution to gene regulation through production of fragmentary peptides and siRNAs, including the well-studied P53 duplications in elephants[8,9].

3 Introduction to linkage modification

RMGDs arise all the time, but the vast majority are often lost. What causes retrogenes to fix? As mentioned above, the process of RMGD implicates the alteration or loss of evolutionary constraints as an important factor in why some retrogenes fix. One constraint that is altered in the process of RMGD is the linkage between the new copy of the gene and its old linkage partners. This brings us to the topic of recombination, and its importance in evolution.

The evolutionary advantage of recombination has long been debated, ever since the publication of R.A. Fisher's influential book, *The Genetical Theory of Natural Selection*[10] in 1930, closely followed by Hermann Muller's paper *Some Genetic Aspects of Sex*[11] in 1932. These two works establish the essential reasoning for why recombination is selected for in evolution - that it both helps advantageous mutations arising in different lineages to simultaneously fix in the population (thereby subverting clonal interference), and allows beneficial mutations on poor genetic backgrounds to be effective selected upon (thereby subverting Hill-Robertson interference).

In 1974, the mathematical arguments were summarized and established to hold under realistic population constraints in Felsenstein's seminal work, *The evolutionary advantage of recombination*[12]. Following this, the outstanding question that remained was what molecular mechanisms could facilitate selection on linkage. Work by Masatoshi Nei in 1967 had previously established mathematical models for the modification of linkage by natural selection, which included 2 mechanisms - one invoking linkage alteration through chromosomal rearrangement, and one relying on a linkage modifying locus in the genome[13]. The latter model received the bulk of attention in years following Felsenstein's publication, with further development by Marcus Feldman[14], Sarah Otto[15,16,17], and Nick Barton[15,17,18]. Their contributions to modifier allele models were critical to establishing that modifier alleles would be capable of increasing and/or maintaining recombination in populations under a variety of environmental conditions.

4 An alternative model for linkage modification

While the theory of modifier alleles has been extensively developed, Nei's model of linkage modification by chromosomal rearrangement has thus far been subject to much less exploration. With our current knowledge of retrogenes and RMGDs, it is potentially interesting, then, to revisit Nei's original model of linkage modification by chromosomal rearrangement through the lens of RMGD. In fact, one aspect of RMGD that has been left relatively unexplored is its interaction with linkage. Genes are often subject to epistasis where different combinations of alleles have non-additive fitness effects. Because RMGD creates new gene copies with different linkage relationships, selection on linkage may influence subsequent fixation as retrogenes. Conversely, fixation as retrogenes may also represent a significant pathway for mediating selection on linkage, where the alternative would be a direct modification of recombination rate. An RMGD-based model of linkage modification would thereby complement the modifier allele models developed by Nei, Feldman, Otto, and Barton.

Nei's original 1967 paper suggests a model of linkage modification by chromosomal rearrangement, but does not suggest a specific molecular mechanism for rearrangement. Implementing his model in light of RMGD potentially overcomes several limitations of modifier allele models. One limitation of modifier allele models is that they lack the ability to modify recombination rates between genes on different chromosomes; RMGD, on the other hand, can bring genes from different chromosomes onto the same chromosome, thereby altering reducing linkage distance. Other potential limitations are structural - can modifier alleles exist for all pairs of linked genes on a chromosome? Can modifier alleles account for all variation in recombination rates? In some areas on the chromosome, like centromeres, it seems unlikely, or at least unprecedented, for increased recombination to arise. Also, mathematically, the conditions that favor increased recombination through modifier allele mutations are limited[15]. Finally, some limitations may be imposed by the presence of other coadapted genes nearby. Increasing recombination between two loci that are separated by one or more coadapted genes would likely disrupt the coadapted complex as well and decrease fitness overall. I propose, then, to investigate the role of selection on linkage in effecting retrogene fixation, and conversely, the role of RMGD in facilitating the movement of genes away from poor genetic backgrounds.

There are a number of empirical results that support the possibility of these roles. One highly investigated area of retrogene evolution is the so-called "out-of-the-X" (ootX) pattern[19]. This is a pattern, documented primarily in mammals and Drosophila, in which we see an excess of retrogenes originating from the X chromosome compared to what we would expect by chance. The expression of ootX retrogenes has been characterized as significantly male-biased, often with specific functions in spermatogenesis; in fact, in mice, mutants for one particular ootX retrogene are rendered completely deficient in spermatogenesis[20]. The most widely accepted hypothesis claims that these retrogenes are favored because they allow for expression during meiotic sex chromosome inactivation[19]. The sex chromosomes, in this case, constitute a poor genetic background in a temporal sense, and RMGD facilitates the movement of genes away from this poor background.

Recent studies also support a more general phenomenon termed retrogene replacement, wherein we see loss of the parental gene coupled with maintenance of its retrogenes. Ciomborowska et al. document 25 such "orphan" retrogenes in humans[7]. Importantly, none of these "orphan" retrogenes exhibit a tendency for testis-specific expression, which sets them apart from the ootX retrogenes and suggests the possibility of a broader selective force at work. Additional examples of retrogene replacement have been found across diverse taxa. Most strikingly, Maeso et al. found that all tetrapods share an instance of retrogene replacement in the dismantling of the ancestral Iroquois-Sowah genomic regulatory block (GRB)[21]. In the ancestral GRB, the regulatory regions for Iroquois lay within the introns of Sowah; however, a retrotransposition event in the tetrapod lineage resulted in the disentanglement of these regulatory constraints. These examples of retrogene replacement show that there exists circumstances where RMGD results in not just the duplication, but the movement of a gene, despite the loss of upstream regulatory regions. Although loss of the parental gene is not strictly necessary for a linkage selective effect to exist (since functional copies may regulate each other to similar effect), the loss does support the possibility of significant interaction between selection on linkage and retrogene fixation.

5 Investigative proposal

I seek to answer two specific questions: (i) does selection on linkage influence retrogene fixation, and (ii) does RMGD help resolve selection on linkage? One can imagine many possible approaches, ranging from the purely quantitative to the experimental. I propose an approach using a combination of theoretical and computational methods, proceeding in three phases: 1) construction of theoretical quantitative models. 2) construction of a general pipeline to identify RMGDs. 3) hypothesis testing.

5.1 Mathematical modeling

Mathematical modeling is an essential component for any investigation in theoretical evolutionary biology. Generally, the idea is to establish reasonable null hypotheses and guide investigations into specific influencing factors. Modeling in this project will attempt to establish a reasonable null for the circumstances under which the emergence of a retrogene will tend to fix, specifically with regards to selection on linkage. This aspect of modeling is being done through stochastic forward-time simulations, wherein changes in gene frequency in a population are tracked across generations. By simulating conditions with selection on linkage, we can examine how often and how long it takes for the population to adjust. One important environmental condition to be tested is where two novel beneficial mutations arise in different lineages within the population. The influence of retrogenes can then be quantified by how long it takes for both mutations to fix in the population when RMGD adjusts the linkage distance between loci; this essentially measures its effect on clonal interference. Another environmental condition to be tested would be where a novel beneficial mutation arises on a poor genetic background. The influence of retrogenes in this scenario can be quantified by the distribution over the alternative genetic backgrounds that an RMGD could transfer that gene to; this would assess effects on Hill-Robertson interference under different conditions. To obtain closed-form estimates, recursion equations from Otto & Barton can be adapted[15].

5.2 Analysis pipeline

The next step would be to construct a data analysis pipeline to collect data for testing hypotheses. This pipeline will use genome and phylogenetic data to identify, sort, and characterize retrogenes. In conjunction with existing databases of retrogenes, such as RetrogeneDB[22], this data will help characterize retrogene content across a wide array of species. This data can then be used for hypothesis testing. This package will be written in R/C++ for maximum flexibility and speed.

5.3 Hypothesis testing

The final aspect of my proposed analysis would be hypothesis testing. The hypotheses to be tested are
(i) does selection on linkage influence retrogene fixation, and
(ii) does RMGD help mediate selection on linkage?
Using known and putative retrogenes identified with the above pipeline, (i) can be tested using data on gene regulatory networks that contain parental genes of retrogenes. These networks represent the set of other genes that the parental gene interacts with, and thus the set of genes most likely to have an epistatic effect on the fitness of different alleles of the parental gene. I will answer (i) by comparing the linkage distances between retrogenes and other members of the same gene regulatory network to a distance matrix for the parental gene. Diverse sets of GRN data are available across various online databases, and the effect of interest can be detected through a Monte Carlo permutation test. To be precise, for both the GRN containing the parental gene only, and the GRN containing the retrogene, a distance matrix will be generated,

$$D(x) = \begin{pmatrix} \delta_{1,1}(x) & \dots & \delta_{1,k}(x) \\ \vdots & \ddots & \vdots \\ \delta_{k,1}(x) & \dots & \delta_{k,k}(x) \end{pmatrix}$$

where $k = number\, of\, genes\, in\, network$ and $\delta = distances\, between\, gene\, and\, network\, neighbor$. The matrix is thus symmetric with 0's along the diagonal. The distance between two of these matrices can then be measured by the L_2 norm, or

$$dist(D(x), D(y)) = \sum_{i=1}^{k} \sum_{j=1}^{k} \|D_{i,k}(x) - D_{i,k}(y)\|_2$$

The null distribution of these distances, given a parental gene, is the distribution resulting from the completely random insertion of the retrogene into the genome, i.e.$X \sim Unif(0, t)$, where 0 and t correspond to an arbitrary "beginning" and "end" of a genome respectively. The null distribution is then defined by the transformation:

$$W \sim dist(D(X), D(y))$$

and a Monte Carlo permutation test can be used to generate an empirical distribution for which our test statistic can be compared.

To answer (ii), I will characterize retrogene formation rates prior to and after major translocation or chromosome fusion events, such as in the wake of the human chromosome 2 fusion. This test would proceed as a simple likelihood-ratio test where retrogene formation within a certain period of time is assumed to be Poisson distributed. We can then test the plausibility of a single rate for all retrogene formation before and after translocation events, or multiple rates. Specifically,

$$\Lambda(x) = \frac{\mathcal{L}(\theta_0|x)}{\mathcal{L}(\hat{\lambda}|x)}$$

where $\mathcal{L}(\theta|x)$ is the log-likelihood of some value x for a $Poisson(\theta)$ distributed random-variable, and $\hat{\theta}$ is the point estimate for the rate parameter based on data immediately following a major translocation.

Significance for these hypotheses tests will be assessed at a Bonferroni-corrected type 1 error rate of .001, which would be sufficiently unexpected as to warrant a closer look. If testing these hypotheses produces significant results, it would constitute significant support for an interaction between selection on linkage and retrogene formation.

6 Conclusion

It is important to note that the results from hypothesis testing are only the beginning. With this research, I hope to not only shed light on the interaction between linkage and retrogenes, but pave the way towards new insights into the evolutionary forces impacting retrogenes, and genome evolution at large. Modifier allele models have been incredibly useful in explaining and accounting for selection on linkage – they are, however, subject to some limitations. By developing a complementary model of RMGD-mediated linkage modification, these limitations are addressed. Gene duplication, overall, is well-studied as a significant driver of species evolution, and my studies will contribute to this body of literature.

References

[1] Henrik Kaessmann, Nicolas Vinckenbosch, and Manyuan Long. Rna-based gene duplication: mechanistic and evolutionary insights. *Nature Reviews Genetics*, 10(1):19–31, 2009. ISSN 1471-0064. doi: 10.1038/nrg2487. URL `https://doi.org/10.1038/nrg2487`.

[2] Claudio Casola and Esther Betrán. The genomic impact of gene retrocopies: What have we learned from comparative genomics, population genomics, and transcriptomic analyses? *Genome Biology and Evolution*, 9(6):1351–1373, 2017. ISSN 1759-6653. doi: 10.1093/gbe/evx081. URL `https://doi.org/10.1093/gbe/evx081`.

[3] David M. Sayah, Elena Sokolskaja, Lionel Berthoux, and Jeremy Luban. Cyclophilin a retrotransposition into trim5 explains owl monkey resistance to hiv-1. *Nature*, 430(6999):569–573, 2004. ISSN 1476-4687. doi: 10.1038/nature02777. URL `https://doi.org/10.1038/nature02777`.

[4] Jianming Zhang, Antony M. Dean, Frédéric Brunet, and Manyuan Long. Evolving protein functional diversity in new genes of drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16246, 2004. doi: 10.1073/pnas.0407066101. URL `http://www.pnas.org/content/101/46/16246.abstract`.

[5] Fabien Burki and Henrik Kaessmann. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nature Genetics*, 36(10):1061–1063, 2004. ISSN 1546-1718. doi: 10.1038/ng1431. URL `https://doi.org/10.1038/ng1431`.

[6] Hongzheng Dai, Ying Chen, Sidi Chen, Qiyan Mao, David Kennedy, Patrick Landback, Adam Eyre-Walker, Wei Du, and Manyuan Long. The evolution of courtship behaviors through the origination of a new gene in drosophila. *Proceedings of the National Academy of Sciences*, 105(21):7478, 2008. doi: 10.1073/pnas.0800693105. URL `http://www.pnas.org/content/105/21/7478.abstract`.

[7] Joanna Ciomborowska, Wojciech Rosikiewicz, Damian Szklarczyk, Wojciech Makałowski, and Izabela Makałowska. "orphan" retrogenes in the human genome. *Molecular Biology and Evolution*, 30(2):384–396, 2013. ISSN 0737-4038. doi: 10.1093/molbev/mss235. URL `https://doi.org/10.1093/molbev/mss235`.

[8] Michael Sulak, Lindsey Fong, Katelyn Mika, Sravanthi Chigurupati, Lisa Yon, Nigel P. Mongan, Richard D. Emes, and Vincent J. Lynch. Tp53 copy number expansion is associated with the evolution of increased body size and an enhanced dna damage response in elephants. *eLife*, 5:e11994, 2016. ISSN 2050-084X. doi: 10.7554/eLife.11994. URL `https://doi.org/10.7554/eLife.11994`.

[9] Lisa M. Abegglen, Aleah F. Caulin, Ashley Chan, Kristy Lee, Rosann Robinson, Michael S. Campbell, Wendy K. Kiso, Dennis L. Schmitt, Peter J. Waddell, Srividya Bhaskara, Shane T. Jensen, Carlo C. Maley, and Joshua D. Schiffman. Potential mechanisms for cancer resistance in elephants and comparative cellular response to dna damage in humans. *JAMA*, 314(17):1850–1860, 2015. ISSN 0098-7484. doi: 10.1001/jama.2015.13134. URL `https://doi.org/10.1001/jama.2015.13134`.

[10] R.A. Fisher. *The genetical theory of natural selection*. Clarendon PRess, Oxford, 1930.

[11] H. J. Muller. Some genetic aspects of sex. *The American Naturalist*, 66(703):118–138, 1932. ISSN 00030147, 15375323. URL `http://www.jstor.org.srv-proxy1.library.tamu.edu/stable/2456922`.

[12] Joseph Felsenstein. The evolutionary advantage of recombination. *Genetics*, 78(2): 737, 1974. URL `http://www.genetics.org/content/78/2/737.abstract`.

[13] M. Nei. Modification of linkage intensity by natural selection. *Genetics*, 57(3):625–641, 1967. ISSN 0016-6731. URL `https://pubmed.ncbi.nlm.nih.gov/5583732 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1211753/`.

[14] Marcus W. Feldman, Freddy B. Christiansen, and Lisa D. Brooks. Evolution of recombination in a constant environment. *Proceedings of the National Academy of Sciences*, 77(8):4838, 1980. doi: 10.1073/pnas.77.8.4838. URL `http://www.pnas.org/content/77/8/4838.abstract`.

[15] S. P. Otto and N. H. Barton. The evolution of recombination: removing the limits to natural selection. *Genetics*, 147(2):879–906, 1997. ISSN 0016-6731. URL `https://pubmed.ncbi.nlm.nih.gov/9335621 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1208206/`.

[16] Sarah P. Otto and Yannis Michalakis. The evolution of recombination in changing environments. *Trends in Ecology & Evolution*, 13(4):145–151, 1998.

ISSN 0169-5347. doi: https://doi.org/10.1016/S0169-5347(97)01260-3. URL http://www.sciencedirect.com/science/article/pii/S0169534797012603.

[17] Sarah P. Otto and Nick H. Barton. Selection for recombination in small populations. *Evolution*, 55(10):1921–1931, 2001. ISSN 0014-3820. doi: 10.1111/j.0014-3820.2001.tb01310.x. URL https://doi.org/10.1111/j.0014-3820.2001.tb01310.x.

[18] N. H. Barton. A general model for the evolution of recombination. *Genetical Research*, 65(2):123–144, 1995. ISSN 0016-6723. doi: 10.1017/S0016672300033140. URL https://www.cambridge.org/core/article/general-model-for-the-evolution-of-recombi

[19] J. J. Emerson, Henrik Kaessmann, Esther Betrán, and Manyuan Long. Extensive gene traffic on the mammalian x chromosome. *Science*, 303(5657): 537–540, 2004. ISSN 0036-8075. doi: 10.1126/science.1090042. URL https://science.sciencemag.org/content/303/5657/537.

[20] Julie Bradley, Andrew Baltus, Helen Skaletsky, Morgan Royce-Tolland, Ken Dewar, and David C. Page. An x-to-autosome retrogene is required for spermatogenesis in mice. *Nature Genetics*, 36(8):872–876, 2004. ISSN 1546-1718. doi: 10.1038/ng1390. URL https://doi.org/10.1038/ng1390.

[21] Ignacio Maeso, Manuel Irimia, Juan J. Tena, Esther González-Pérez, David Tran, Vydianathan Ravi, Byrappa Venkatesh, Sonsoles Campuzano, José Luis Gómez-Skarmeta, and Jordi Garcia-Fernàndez. An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement. *Genome Research*, 22(4):642–655, 2012. doi: 10.1101/gr.132233.111. URL http://genome.cshlp.org/content/22/4/642.abstract.

[22] Michał Kabza, Joanna Ciomborowska, and Izabela Makałowska. Retrogenedb–a database of animal retrogenes. *Molecular biology and evolution*, 31 (7):1646–1648, 2014. ISSN 1537-1719 0737-4038. doi: 10.1093/molbev/msu139. URL https://pubmed.ncbi.nlm.nih.gov/24739306 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4069623/.